dx.doi.org/10.22093/wwj.2021.307401.3182

Original Paper

Journal of Water and Wastewater, Vol. 33, No. 2, pp: 89-106

Identification and Prediction of Consumption Behavior Using Decision Tree and Consumer Value Pyramid

A. Mozafari[1], A. R. Ali Ahmadi[2], A. Mozafari[3]

1. MSc., Dept. of Industrial Engineering, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
(Corresponding Author) mozafaria@abfa-shiraz.ir
2. Prof., Dept. of Information Technology, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
3. MSc., Dept. of Industrial Engineering, Faculty of Industrial Engineering, Payame Noor University, Tehran, Iran

(Received Sep. 30, 2021 Accepted Feb. 8, 2022)

To cite this article:
Mozafari, A., Ali Ahmadi, A. R., Mozafari, A. 2022. "Identification and prediction of consumption behavior using decision tree and consumer value pyramid" Journal of Water and Wastewater, 33(2), 89-106.
Doi: 10.22093/wwj.2021.307401.3182. (In Persian)

Abstract

What is important in the current situation it is important to pay attention to patterns of the water consumption behavior of subscribers and to identify subscribers and consumers with a higher position in the value pyramid in the field of water consumption management policy. Value pyramid is a tool that identifies valuable customers in terms of consumption, and so, has great application and importance in the field of water consumption management to identify high-consumption and low-consumption customers. Therefore, in the present study, in order to identify the patterns of consumption behavior of Shiraz Water and Wastewater Company subscribers based on their consumption value pyramid and to predict the subscribers with a higher position in the value pyramid, data mining techniques have been used. In the framework of the proposed method, first, the data of water subscribers' consumption including residential, commercial and industrial subscribers, public and administrative, religious and educational sites, military and non-governmental, for two consecutive years, were extracted from the database of Shiraz Water and Wastewater Company. After determining the optimal number of clusters using self-organizing map and Davis Bouldin index, clustering operation is performed using K-Means algorithm. It should be noted that the indicators and criteria for subscriber clustering include the type of use, location, consumption, history of unauthorized branching, number of disconnection notices and time of payment of bills which have been identified using

the opinion of experts. Then, while calculating the consumption value of subscribers in each cluster and plotting the subscriber consumption value pyramid, the decision tree algorithm is used to predict and discover the behavioral patterns of subscribers. The results show that Shiraz Water and Wastewater Company subscribers are divided into six clusters in terms of consumption behavior patterns. While drawing the consumption value pyramid, these six clusters are classified into three classes: high consumption subscribers, medium consumption subscribers and low consumption subscribers. After implementing the decision tree, the accuracy of the tree was 78.92 that, according to the results of the decision tree, the subscribers of these three classes have 11 patterns of behavior that predict the type of consumption. Thus, according to the 11 behavioral patterns of the subscribers of Shiraz Water and Wastewater Company, the consumption of new subscribers can be predicted and its position in the value pyramid can be determined.

Keywords: K-Means Algorithm, Data Mining, Self-Organized Neural Network, Shiraz Water and Wastewater Company, Water Consumption Management, Water Subscribers.

مجله آب و فاضلاب، دوره ۳۳، شماره ۲، صفحه: ۱۰۶-۸۹

شناسایی و پیش‌بینی رفتار مصرف با استفاده از درخت تصمیم و هرم ارزش مصرف مشترکین

عظیمه مظفری[۱]، علیرضا علی احمدی[۲]، اعظم مظفری[۳]

۱- کارشناسی ارشد، گروه مهندسی صنایع، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران
(نویسنده مسئول) mozafaria@abfa-shiraz.ir
۲- استاد، گروه فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران
۳- کارشناسی ارشد، گروه مهندسی صنایع، دانشکده مهندسی صنایع، دانشگاه پیام‌نور، تهران، ایران

(دریافت ۱۴۰۰/۷/۸ پذیرش ۱۴۰۰/۱۱/۱۹)

برای ارجاع به این مقاله به صورت زیر اقدام بفرمایید:
مظفری، ع.، علی احمدی، ع. ر.، مظفری، ا.، ۱۴۰۱، " شناسایی و پیش‌بینی رفتار مصرف با استفاده از درخت تصمیم و هرم ارزش مصرف مشترکین "
مجله آب و فاضلاب، ۳۳(۲)، ۱۰۶-۸۹. **Doi: 10.22093/wwj.2021.307401.3182**

چکیده

آنچه در شرایط کنونی اهمیت دارد، توجه به الگوی رفتار مصرف آب و شناخت مشترکین و مصرف‌کنندگان با جایگاه بالاتر در هرم ارزش در حوزه سیاست‌گذاری مدیریت مصرف آب است. هرم ارزش، ابزاری است که مشترکین باارزش از نظر مصرف را شناسایی می‌کند، بنابراین کاربرد و اهمیت زیادی در حوزه مدیریت مصرف آب به‌منظور شناسایی مشترکین پرمصرف و کم‌مصرف دارد. ازاین‌رو در این پژوهش به‌منظور شناسایی الگوی رفتاری مصرف مشترکین شرکت آب و فاضلاب شیراز بر مبنای هرم ارزش مصرف آنها و پیش‌بینی مشترکین با جایگاه بالاتر در هرم ارزش، از تکنیک‌های داده‌کاوی استفاده شد. در چارچوب روش پیشنهادی ابتدا داده‌های مربوط به مصرف مشترکین آب شامل مشترکین مسکونی، تجاری و صنعتی، عمومی و اداری، اماکن مذهبی و آموزشی، نظامی و غیردولتی برای ۲ سال متوالی از پایگاه داده شرکت آب و فاضلاب شیراز استخراج شد و پس از تعیین تعداد بهینه خوشه با استفاده از شبکه عصبی خودسازمان‌ده و شاخص دیویس بولدین، عملیات خوشه‌بندی با استفاده از الگوریتم Kمیانگین انجام شد. لازم به‌ذکر است که شاخص‌ها و معیارهای خوشه‌بندی مشترکین شامل نوع کاربری، محل سکونت، میزان مصرف، سابقه انشعاب غیرمجاز، تعداد اخطار قطع و زمان پرداخت قبوض هستند که با استفاده از نظر متخصصین مشخص شده‌اند. در ادامه ضمن محاسبه ارزش مصرف مشترکین هر خوشه و ترسیم هرم ارزش مصرف مشترکین، با استفاده از الگوریتم درخت تصمیم به پیش‌بینی و کشف الگوهای رفتاری مشترکین پرداخته شد. نتایج نشان داد که مشترکین شرکت آب و فاضلاب شیراز از نظر الگوی رفتار مصرف در ۶ خوشه قرار می‌گیرند که ضمن ترسیم هرم ارزش مصرف، این ۶ خوشه در ۳ کلاس مشترکین با مصرف زیاد، مشترکین با مصرف متوسط و مشترکین کم‌مصرف دسته‌بندی شدند. پس از پیاده‌سازی درخت تصمیم، صحت درخت برابر با ۷۸/۹۲ بود که بر اساس آن مشترکین این ۳ کلاس، ۱۱ الگوی رفتاری داشتند که پیش‌بینی کننده نوع مصرف بود. به‌این ترتیب طبق ۱۱ الگوی رفتاری مصرف مشترکین شرکت آب و فاضلاب شیراز می‌توان میزان مصرف مشترکین جدید را پیش‌بینی و جایگاه آن در هرم ارزش را تعیین کرد.

واژه‌های کلیدی: الگوریتم Kمیانگین، داده‌کاوی، شبکه عصبی خودسازمان‌ده، شرکت آب و فاضلاب شیراز، مدیریت مصرف آب، مشترکین آب

dx.doi.org/10.22093/wwj.2021.307401.3182

۱- مقدمه

آب به‌عنوان یک کالای اقتصادی و اجتماعی است که حیات همه موجودات زنده به آن وابسته بوده و هر چند که از منابع تجدیدشونده به‌شمار می‌رود ولی مقدار آن محدود است (Taherdoost, 2021). در دسترس بودن منابع انرژی مانند آب، نقش مهمی در رشد توسعه اقتصادی، اجتماعی و سیاسی هر جامعه‌ای داشته است (Oyedepo, 2014). بنابراین، مدیریت منابع آب و اصلاح الگوی مصرف بسیار حائز اهمیت بوده و نقشی کلیدی در حوزه امنیت آب، امنیت انرژی، حفظ محیط‌زیست و همچنین تصمیم‌گیری‌های کلان اقتصادی و سیاسی دارد (Malekmohamadi and Mozafari, 2018).

بر اساس گزارش مؤسسه جهانی منابع[1]، ایران از لحاظ شاخص منابع آب، در دسته آسیب‌پذیر قرار داشته و بخش‌های زیادی از آن دچار کم‌آبی است. به‌طور کلی متوسط نزولات آسمانی ایران حدود ۲۶۲ میلی‌متر در سال است. ۶۵ درصد کشور ما را مناطق خشک و نیمه‌خشک تشکیل می‌دهند که به‌طور متوسط مقدار بارندگی در آنها از ۱۵۰ میلی‌متر کمتر است. در بیشتر این مناطق تنها منبع آبی تأمین‌کننده تقاضای بخش‌های اقتصادی و اجتماعی، ذخایر آبی آبخوان‌ها هستند که در چند دهه اخیر به‌علت مازاد برداشت‌ها با بیلان منفی روبه‌رو مواجه هستند (Noori et al., 2015).

مصرف آب در ایران بیش از مصرف سرانه آب در مکان‌هایی است که از نظر آب و هوایی و زندگی اجتماعی و اقتصادی مشابه ایران هستند که این موضوع منجر به بروز بحران کم‌آبی شده و نیازمند توجه ویژه به مدیریت مصرف آب است (Foster and Beattie, 1988). بنابراین با توجه به شرایط کنونی بحران آب و افزایش تقاضا برای این ماده حیاتی و کالای ارزشمند، نگاه فراگیر به آب و مدیریت آن اهمیت خاصی دارد (Amini et al., 2018a).

همان طور که با بررسی پژوهش‌های پیشین مشاهده می‌شود، افزایش روزافزون تقاضا برای مصارف مختلف آب و تأثیر عوامل مختلف بر آن، همچنین لزوم مدیریت مصرف توسط مشترکین ایجاب می‌کند که با برطرف کردن مسائلی که در حوزه مصرف آب وجود دارد، نسبت به کاهش مصرف بی‌رویه آب اقدام کرد (Ansari et al., 2015).

برای اجرای موفقیت‌آمیز مدیریت مصرف آب، لازم است رفتار مشترکین یا مصرف‌کنندگان بررسی و در صورت لزوم اصلاح شود. ازاین‌رو امروزه موضع آب و بررسی الگوها و نحوه مصرف آن، اهمیت زیادی برای شرکت‌های آب و فاضلاب دارد و ارائه هر گونه راهکاری برای کاهش مصرف و جلوگیری از هدررفت آن بسیار باارزش است. شرکت‌های آب و فاضلاب همواره برای جلب رضایت مشترکین، تأمین و توزیع آب با قیمت مناسب و حداقل کردن هزینه‌ها تلاش کرده‌اند. این اهداف هنگامی میسر خواهد شد که این شرکت‌ها شناخت و درک صحیحی نسبت به رفتار مصرفی مشترکین داشته باشند (Kojury Naft Chali and Fereydonian, 2015).

با توجه به لزوم و اهمیت بررسی الگوی رفتاری مصرف‌کنندگان آب در شرایط کم‌آبی دنیای امروزی، پژوهش‌های بسیاری در زمینه بررسی و شناسایی عوامل مؤثر بر الگوی رفتاری مصرف آب انجام شده است، ازجمله (Maleki Nasab et al., 2007) به ارزیابی صرفه‌جویی در مصرف آب خانگی به واسطه استفاده از قطعات کاهنده مصرف پرداخته است، (Garcia et al., 2013)مشخصات اجتماعی و جمعیتی در پیامدهای نگرش ها و رفتارهای مربوط به آب در امتداد ساحل مدیترانه را بررسی کرده‌اند، (Willis et al., 2013) تأثیر عوامل اجتماعی و جمعیتی و دستگاه‌های کارآمد را بر مصرف نهایی آب در خانوارها بررسی کرده‌اند، (Rathnayaka et al., 2014) عوامل مؤثر بر تغییرپذیری مصرف آب خانگی در ملبورن را ارزیابی کرده‌اند،(Mohammadi, 2014) تأثیر هدفمندی یارانه‌ها بر الگوی مصرف و میزان مصرف آب شهر اردبیل را بررسی کرده‌اند، (Chen et al., 2015)یک مدل معیار برای مصرف آب خانگی بر اساس شبکه‌های منطقی تطبیقی ارائه کرده‌اند، (Castellano, 2020) وضعیت حفاظت از حقوق آب و استفاده غیرمجاز از آب در غرب آمریکا را بررسی کرده است.

در برخی از پژوهش‌ها نیز با استفاده از روش‌ها و تکنیک‌های مختلف، مصرف آب پیش‌بینی شده که از جمله می‌توان به پژوهش (Shamsai, 2000) اشاره کرد که تابع تقاضای آب استان اصفهان را برآورد و پیش‌بینی کرده است، (Ebrahimi and Naderi, 2001) مدیریت عرضه و تقاضای آب شرب در شرایط خشک‌سالی اصفهان را بررسی و ارزیابی کرده است،(Sabouhi and Noubakht, 2009) تقاضای آب شهر پردیس را برآورد و پیش‌بینی کرده‌اند،

[1] World Resources Institute

(Javadianzade, 2009) تابع تقاضای آب شهری را با استفاده از روش شبکه‌های عصبی مصنوعی پیش‌بینی کرده است، (Tabesh and Dini, 2010) تقاضای روزانه آب شهری را با استفاده از شبکه‌های عصبی مصنوعی پیش‌بینی کرده‌اند، (Aghahosseinali Shirazi and Akbarpour, 2011) تقاضای روزانه آب شهری را با استفاده از سری فوریه برآورد کرده‌اند.

همان طور که مشاهده می‌شود در این پژوهش‌ها معمولاً از روش‌های آماری، سری‌های زمانی و شبکه عصبی استفاده شده، در حالی‌که تکنیک‌های داده‌کاوی[1] نیز به‌عنوان یکی از ابزارهای تحلیلی برای کشف الگوها و روابط بین داده‌ها در سالیان اخیر در پژوهش‌های بسیاری استفاده شده که دقت و صحت بالایی دارد (Berry and Linoff, 2004).

مزیت استفاده از روش‌های داده‌کاوی برای آن قابل بحث است که ابزارهای مفیدی به‌منظور استخراج اطلاعات از پایگاه‌های داده بزرگ با حجم انبوهی از داده‌ها هستند که از طریق آن داده‌ها می‌توانند به‌صورت مؤثر ذخیره و استخراج شوند (Monika and Amarpreet, 2018).

شرکت‌های آب و فاضلاب داده‌های زیادی راجع به مشترکین خود ذخیره و نگهداری می‌کنند و این درحالی است که در کشف دانش یا ارزش نهفته در این داده‌ها کم‌توان هستند. پایگاه داده مشترکین یکی از مهم‌ترین دارایی‌های این‌گونه شرکت‌ها است که می‌توانند از آن برای توسعه استراتژی در درک بهتر الگوی رفتاری مشترکین و مصرف‌کنندگان با استفاده از تجزیه‌وتحلیل مشخصات و الگوی رفتاری مصرف هر مشترک استفاده کنند. با توجه به آنچه در مورد تکنیک‌های داده‌کاوی گفته شد و ضمن در نظر گرفتن گسترش فناوری اطلاعات و ارتباطات در سازمان‌ها، با به‌کارگیری بانک‌های اطلاعاتی موجود و استفاده از ابزارها و الگوریتم‌های مفیدی مانند داده‌کاوی، می‌توان ماهیت پیچیده داده‌های مرتبط با الگوی رفتار مصرفی مشترکین و روابط نامحسوس میان این داده‌ها را مدل کرده و الگوهای مصرف را شناسایی کرد (Amini et al., 2018a).

با استفاده از الگوریتم‌های داده‌کاوی می‌توان ماهیت پیچیده رفتار مصرفی مشترکین را در قالب الگوهایی شناسایی کرد که کمک شایسته‌ای به مدیریت مصرف می‌کند (Amini, 2020).

از جمله پژوهش‌هایی که از تکنیک‌های داده‌کاوی در صنعت آب و فاضلاب استفاده کرده‌اند نیز می‌توان به پژوهش‌های (Díaz et al., 2010) اشاره کرد که وظایف قبل و بعد از پردازش در داده‌کاوی برای یک مشکل دنیای واقعی را بررسی کرده‌اند، (Boyle et al., 2011) داده‌های صورت‌حساب آب برای اطلاع از سیاست‌ها و استراتژی‌های ارتباطی را بررسی کرده‌اند، (Aghababaei et al., 2011) یک مدل داده‌کاوی برای مشترکین بدهکار شرکت آب و فاضلاب مشهد طراحی کرده‌اند، (Yurekli et al., 2012) خشکسالی فصلی و سالانه منطقه‌ای را با استفاده از روش داده‌کاوی پیش‌بینی کرده‌اند، (Aghababaei and Shahrabi, 2012) از داده‌کاوی در مواجهه با مشترکین بدهکار در شرکت آب و فاضلاب مشهد استفاده کرده‌اند، (Dutta and Chaki, 2012) کاربردهای داده کاوی در مدیریت کیفیت آب را بررسی کرده است، (Khan et al., 2012) به ارزیابی چندین روش داده‌کاوی برای پیش‌بینی نیاز آبیاری پرداخته‌اند، (Hashem, 2012) یک مدل تشخیص کلاهبرداری مبتنی بر داده‌کاوی برای سیستم صورت‌حساب مصرف آب ارائه کرد، (Wen et al., 2013) سیستم تجزیه‌وتحلیل مصرف آب را بر اساس داده‌کاوی طراحی کرده‌اند، (Azhar et al., 2013) به بهینه‌سازی نظارت بر کیفیت آب بر اساس الگوریتم‌های فازی پرداخته‌اند، (Mohammad Taheri, 2013) کیفیت پساب خروجی تصفیه‌خانه فاضلاب را با استفاده از داده‌کاوی پیش‌بینی کرد، (Arora et al., 2014) از تجزیه‌وتحلیل خوشه‌ای برای بهبود نظارت بر کیفیت آب رودخانه ساتلوج استفاده کرده‌اند، (Gu et al., 2014) عوامل مؤثر بر کیفیت آب برای مخازن آب آشامیدنی را شناسایی و ارزیابی کرده‌اند، (Sattari and Rezazadeh Judi, 2018) کیفیت آبهای سطحی را با استفاده از روش درخت تصمیم پیش‌بینی کرده‌اند، (Monedero et al., 2015) روشی برای تشخیص دستکاری در کنتورهای آب ارائه کرده‌اند، (Azimi et al., 2015) به ارزیابی و برنامه‌ریزی در برآورد دبی‌های روزانه رودخانه لیقوان پرداخته‌اند، (Soleimanpour et al., 2015) شاخص‌های مؤثر بر کیفیت آب آشامیدنی را با استفاده از تکنیک داده‌کاوی بررسی کرده‌اند، (Jahanpour and Nosrati, 2015) مروری بر کاربردهای داده‌کاوی در مدیریت شرکت آب و فاضلاب داشته‌اند، (Rayegan Shirazinejad et al., 2015)

[1] Data Mining

dx.doi.org/10.22093/wwj.2021.307401.3182

مدل‌های آماری در مدل‌سازی فرایند تصفیه فاضلاب را با استفاده از روش داده‌کاوی بررسی کرده‌اند، (Kazemi, 2015) فرایند داده‌کاوی را برای بهبود فرایندهای مدیریت دانش در مراکز تماس ۱۲۲ به‌کار گرفت،(Thompson, 2016) توصیه‌های آب آشامیدنی را در جوامع ملل اول از طریق داده‌کاوی بررسی کرد، (Cho, 2016) یک مدل ارزیابی کیفیت آبخیز با استفاده از داده کاوی ارائه کرد، (Ji et al., 2016) از تکنیک‌های داده کاوی در عملیات تأمین آب استفاده کرده‌اند، (Soleimanpour et al., 2016) الگوریتم‌های داده‌کاوی بخش‌بندی و درخت تصمیم را برای تعیین مؤثرترین عوامل کیفیت آب آشامیدنی استفاده کرده‌اند، (Amini et al., 2018a) به‌منظور شناسایی مشترکین با مصارف غیرمجاز آب از تکنیک‌های داده‌کاوی استفاده کرده‌اند، (Amini et al., 2018b) به‌منظور استخراج الگوی مصرف آب از تکنیک‌های داده‌کاوی استفاده کرده‌اند، (Khalfi et al., 2018) الگوی مصرف آب خانگی را با رویکرد بخش‌بندی مصرف‌کنندگان بررسی کرده‌اند، (Sattari et al., 2014) با استفاده از روش‌های داده‌کاوی به مدل‌سازی رواناب ماهانه پرداخته‌اند، (Soleimanpour et al., 2018) از تکنیک‌های داده کاوی درخت تصمیم در تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی استفاده کرده‌اند، (Ahangarkani and Khasteh, 2019) مصرف آب شهری (خانگی) شهرستان بابل را با استفاده از روش‌های داده‌کاوی تحلیل کرده‌اند، (Eskandary et al., 2020) شاخص‌های مشارکت عمومی- خصوصی در صنعت آب و فاضلاب ایران را از طریق الگوریتم‌های داده‌کاوی شناسایی کرده‌اند، (Amini, 2020) به مدل‌سازی تشخیص مصرف غیرمجاز آب با استفاده از داده‌کاوی پرداخته است. همان طور که مشاهده شد تکنیک‌های داده‌کاوی در صنعت آب و فاضلاب کاربردهای گوناگونی داشته‌اند.

با توجه به اهمیت این موضوع، این پژوهش با هدف شناسایی و پیش‌بینی الگوی رفتاری مصرف مشترکین شرکت آب و فاضلاب شیراز انجام شد که برای این منظور از تکنیک‌های داده‌کاوی اعم از خوشه‌بندی و درخت تصمیم در کنار مفهوم مهمی به نام هرم ارزش مصرف مشترکین استفاده شده که تاکنون پژوهشی در این راستا و در صنعت آب و فاضلاب انجام نشده است. با استفاده از روش پیشنهادی می‌توان الگوی مصرف را استخراج کرد و مشترکین کم‌مصرف و با جایگاه بالاتر در هرم ارزش را بر اساس هرم ارزش مصرف شناسایی و با استفاده از درخت تصمیم پیش‌بینی کرد.

۲- روش پژوهش

شرکت آب و فاضلاب شیراز با هدف تأمین پایدار نیازهای پایه آب شرب و بهداشتی و همچنین جمع آوری، انتقال، تصفیه و دفع بهداشتی فاضلاب تأسیس شده که تا پایان سال ۱۳۹۹ حدود ۷۵۵ هزار مشترک آب را تحت پوشش داده است. مشترکین شرکت آب و فاضلاب شیراز در ۵ منطقه شهری سکونت دارند، به‌طوری که منطقه ۱ شامل بخش وسیعی از مشترکین مسکونی در حاشیه شرقی شهرستان می‌شود، منطقه ۲ معمولاً شامل مشترکین مناطق مرکزی شهرستان بوده که تا حدود زیادی نیز تجاری هستند، منطقه ۳ عمده مشترکین مسکونی در قسمت شمالی شهرستان شیراز را پوشش می‌دهد، منطقه ۴ نیز شامل مشترکین غرب و جنوب شیراز بوده و منطقه ۵ مشترکین شهر جدید صدرا را در برمی‌گیرد که در فاصله ۲۰ کیلومتری شهرستان شیراز واقع شده است. با توجه به گستره فعالیت این شرکت در سطح شهرستان شیراز و ضمن بررسی روند مصرف آب، در این پژوهش با استفاده از تکنیک‌های داده‌کاوی به شناسایی و پیش‌بینی الگوی رفتاری مصرف مشترکین آب شامل مشترکین مسکونی، تجاری و صنعتی، عمومی و اداری، اماکن مذهبی و آموزشی، نظامی و مشترکین غیردولتی شهرستان شیراز پرداخته شد. لازم به‌ذکر است که برای این منظور از مفهوم هرم ارزش مصرف مشترکین نیز استفاده شد.

یکی از ابزارهای مفیدی که به‌منظور استخراج اطلاعات از این پایگاه‌های داده و نیاز کاربران استفاده می‌شود، داده‌کاوی است که از طریق آن داده‌ها می‌توانند به‌صورت مؤثر ذخیره و استخراج شوند (Monika and Amarpreet, 2018). داده‌کاوی فرایندی است که از ابزارهای تحلیلی گوناگونی برای کشف الگوها و روابط بین داده‌ها استفاده می‌کند (Berry and Linoff, 2004). ازاین‌رو فرایندی است کاملاً علمی برای کشف دانش‌های پنهان و روابط ناشناخته بین داده‌ها که با نگرشی نو، به مسئله استخراج داده‌ها می‌پردازد (Tabatabai, 2009) که در این پژوهش استفاده شد و برای پیاده‌سازی آن، نرم‌افزار SPSS Modeler به‌کار گرفته شد.

این پژوهش به‌دلیل انجام در یک مجموعه مشخص (شرکت آب و فاضلاب شیراز) بر اساس طبقه‌بندی بر مبنای هدف، از نوع پژوهش‌های کاربردی است. زیرا نتایج حاصل از آن می‌تواند

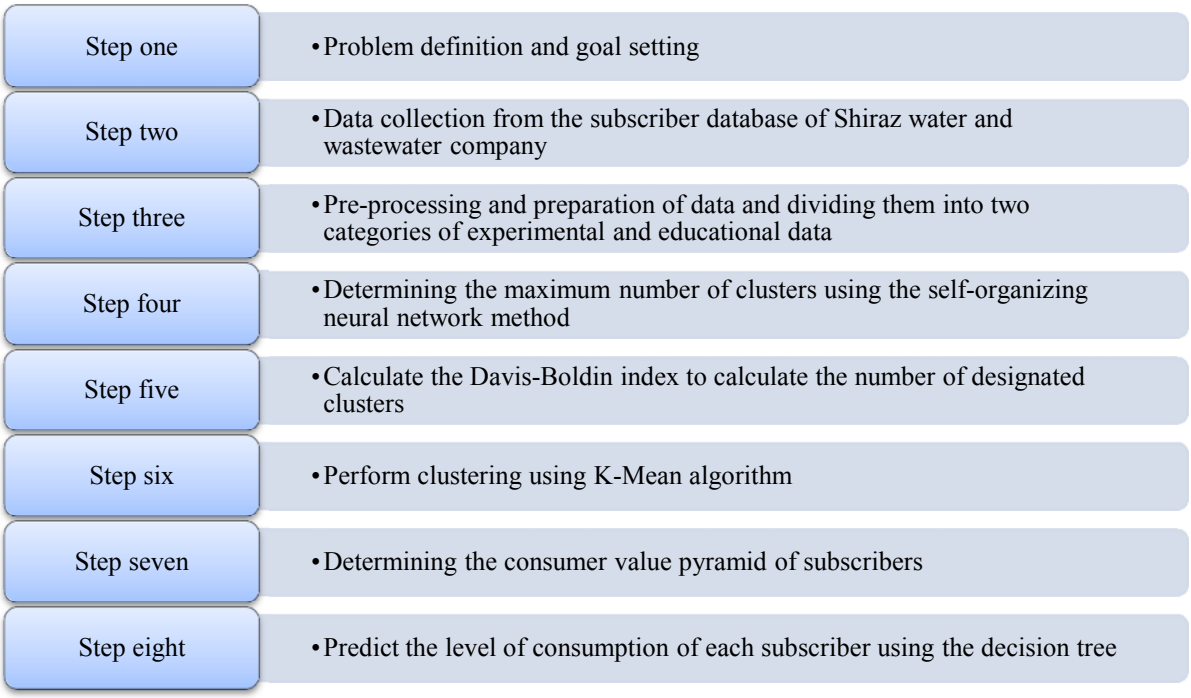

Fig. 1. Research steps

شکل ۱- گام‌های انجام پژوهش

مورد استفاده سازمان‌ها و شرکت‌های آب و فاضلاب در اتخاذ سیاست‌های مناسب قرار گیرد. مراحل انجام پژوهش برای کشف دانش در دستیابی به هدف موردنظر در شکل ۱ نشان داده شده است.

گام اول: بیان مسأله و تعیین هدف

هدف اصلی این پژوهش شناسایی و پیش‌بینی الگوی رفتاری مصرف مشترکین آب شهرستان شیراز با تکیه بر شناخت ارزش مشترکین از نظر نوع مصرف بود.

گام دوم: جمع آوری داده از پایگاه داده مشترکین شرکت آب و فاضلاب شیراز

یکی از مهم‌ترین مراحل فرایند استخراج دانش، جمع‌آوری داده‌های خام و شناسایی اطلاعات مفید برای حل مسئله است. پس از شناخت مسئله باید در پی شناسایی اطلاعات مفید برای حل مسئله و انتخاب منابع داده‌ی مناسب بود. به‌طور کلی برای پیاده‌سازی تکنیک‌های داده‌کاوی باید معیارهایی در نظر گرفته شوند که بتوان بر اساس آنها مشترکین را به گروه‌های مختلف خوشه‌بندی کرده و مصرف مشترکین هر خوشه را تعیین کرد. داده‌های استفاده شده در این پژوهش مربوط به مشترکین شرکت آب و فاضلاب در سال‌های ۱۳۹۸ و ۱۳۹۹ بوده که از پایگاه داده آن استخراج شد.

گام سوم: پیش‌پردازش و آماده‌سازی داده‌ها و تقسیم آنها به دو دسته داده‌های آزمایشی و آموزشی

پس از اینکه داده‌های موردنیاز برای خوشه‌بندی مشترکین، از پایگاه داده شرکت آب و فاضلاب شیراز جمع‌آوری شدند، باید برای استفاده از الگوریتم‌های داده‌کاوی آماده شوند. به‌وسیله تکنیک‌های آماده‌سازی داده می‌توان کیفیت داده‌ها و در نتیجه کیفیت نتایج خروجی را افزایش داد. آماده‌سازی داده‌ها که تحت عنوان پیش‌پردازش شناخته می‌شود، بیشترین زمان را به خود اختصاص می‌دهد و طبق ارزیابی‌های انجام شده، یکی از مهم‌ترین مراحل کشف دانش است، به‌طوری که ۷۰ درصد زمان و ۸۰ درصد موفقیت کل فرایند به این بخش اختصاص می‌یابد. آماده‌سازی داده‌ها شامل عملیاتی چون پاک‌سازی، حذف رکوردهای با تعداد زیاد داده مفقودی، تبدیل ویژگی‌ها، نرمال‌سازی، گسسته‌سازی،

dx.doi.org/10.22093/wwj.2021.307401.3182

حذف داده‌های پرت و ایجاد ویژگی‌های جدید است.

اولین قدم در آماده‌سازی داده‌ها، انتخاب ویژگی‌های مناسب و کاهش داده‌ها است، بنابراین برای بهبود کیفیت داده‌ها، داده‌های مفقود از پایگاه داده حذف شدند؛ همچنین برخی از تبدیل ویژگی‌ها برای استخراج ویژگی‌های مناسب از روی داده‌های موجود انجام شد، به‌این ترتیب در پایان، اطلاعات مربوط به ۵۵۷ هزار مشترک آب شهرستان شیراز باقی ماند.

گام چهارم: تعیین حداکثر تعداد خوشه‌ها با استفاده از روش شبکه عصبی خودسازمان‌ده

از جمله تکنیک‌هایی که برای تعیین خوشه بهینه در فرایند خوشه‌بندی[1] استفاده می‌شود، شبکه عصبی مصنوعی[2] خودسازمان‌ده[3] است که بر پایه اتصال به هم پیوسته چندین واحد پردازشی به نام نرون[4]، ساخته می‌شود که در لایه‌های مشخصی قرار گرفته‌اند (Rosenblatt, 1962). مزایای این روش از جمله تحمل زیاد داده‌های مغشوش[5]، کارکرد موازی، قابلیت استفاده زمانی که دانش بسیار کمی در مورد مسئله داریم، استفاده از آن را در بسیاری از کاربردها افزایش یافته است. برای پیاده‌سازی شبکه عصبی خودسازمان‌ده در این پژوهش دو لایه برای ورودی و خروجی در نظر گرفته شد که لایه اول ۱ نرون و لایه دوم شامل ۶ نرون بود.

گام پنجم: محاسبه شاخص دیویس- بولدین[6] برای محاسبه تعداد خوشه‌های تعیین شده

نتایج حاصل از الگوریتم‌های خوشه‌بندی روی یک مجموعه داده با توجه به انتخاب‌های پارامترهای الگوریتم‌ها می‌تواند بسیار متفاوت از یکدیگر باشد. دو معیار پایه برای انتخاب خوشه‌های بهینه عبارت‌اند از: تراکم (داده‌های متعلق به یک خوشه باید تا حد ممکن به یکدیگر نزدیک باشند) و جدایی (خوشه‌ها خود باید به اندازه کافی از یکدیگر جدا باشند) (Kovács et al., 2013).

یکی از شاخص‌های پرکاربرد برای تعیین تعداد بهینه خوشه، شاخص دویس- بولدین است از شباهت بین دو خوشه استفاده می‌کند که بر اساس پراکندگی یک خوشه و عدم شباهت بین دو خوشه تعریف می‌شود. این شاخص میانگین شباهت بین هر خوشه با شبیه‌ترین خوشه به آن را محاسبه می‌کند و هرچه مقدار آن بیشتر باشد، خوشه‌های بهتری تولید شده است (Davidson, 2002).

گام ششم: انجام خوشه‌بندی با استفاده از الگوریتم Kمیانگین[7]

یکی از کاربردی‌ترین تکنیک‌های داده‌کاوی، الگوریتم خوشه‌بندی است که به‌منظور استخراج و شناسایی الگو از بین حجم انبوه داده‌ها استفاده می‌شود. این تکنیک داده‌های مشابه را در یک خوشه قرار داده و الگویی را به‌عنوان نماینده ارائه می‌کند، به‌طوری که این نماینده بیانگر رفتار داده‌های آن خوشه است (Amoozegar, 2016).

در بحث خوشه‌بندی، هدف این است که داده‌هایی که در یک خوشه قرار دارند بیشترین تشابه را با یکدیگر و کمترین تشابه را با اعضای خوشه‌های دیگر داشته باشند (Kojury Naft Chali and Fereydonian, 2015). یکی از الگوریتم‌های خوشه‌بندی، الگوریتم Kمیانگین است که با توجه به تعداد خوشه‌ها، دسته‌بندی را آغاز کرده و با استفاده از تکنیک جابه‌جایی تکراری، تلاش می‌کند که با جابه‌جایی اشیا از یک خوشه به خوشه دیگر، دسته‌بندی را بهبود دهد. این بهبود به‌طور معمول با کمینه‌سازی یک تابع هدف تعریف شده میسر می‌شود. الگوریتم خوشه‌بندی Kمیانگین (MacQueen, 1967) از دو مرحله اصلی تشکیل شده است که در مرحله اول داده‌ها به نزدیک‌ترین خوشه نسبت داده می‌شوند و در مرحله دوم مرکز خوشه‌ها با توجه به داده‌هایی که اختصاص یافته‌اند دوباره محاسبه می‌شوند (Davies and Bouldin, 1979).

یکی از مشکلات اساسی روش Kمیانگین این است که تعداد خوشه اولیه را تشخیص نداده و به‌منظور پیاده‌سازی آن لازم است که با روش دیگری، تعداد خوشه اولیه تعیین شود.

گام هفتم: تعیین هرم ارزش مصرف مشترکین

برای تعیین مصرف مشترکین در هر خوشه بر مبنای هرم ارزش،

[1] Clustring
[2] Artificial Neural Network
[3] Self Organization Map (SOM)
[4] Neuron
[5] Noisy Data
[6] Davis-Bouldin
[7] K-Means

جدول ۱- معرفی شاخص‌های پژوهش

Table 1. Research indicators

Row	Field	Unit	Data type	Indicator type	Indicator code	Scope of change
1	Type of application	---	Nominal	---	F1	Residential, commercial and industrial, public and administrative, religious and educational places, military, non-governmental
2	Address	---	Nominal	---	F2	Region 1-5
3	Consumption	Cubic meters	Numerical	Reverse	F3	0-126,006,482
4	History of unauthorized branching	---	Nominal	---	F4	Yes- No
5	Interrupt warning number	Once	Numerical	Reverse	F5	0-10
6	Time to pay bills	Day	Numerical	Reverse	F6	

ابتدا شاخص‌های مصرف مشترکین شامل میزان مصرف، سابقه انشعاب غیرمجاز، تعداد اخطار قطع و زمان پرداخت قبوض نرمال می‌شود. با توجه به معکوس بودن شاخص‌های مصرف مشترکین، نرمال‌سازی به شرح زیر انجام می‌شود

$$v^1 = \frac{maxA - v}{maxA - minA} \quad (۱)$$

که در آن

v مقداری است که قرار است نرمال شود، minA و maxA به‌ترتیب کمترین و بیشترین مقدار بین مقدار شاخص A هستند. سپس میانگین مجموع آنها برای مشترکین هر خوشه محاسبه می‌شود.

در این هرم، به ارزش دوره عمر مشترکین اشاره شده است. هر چه از سطح پایین این هرم به سطوح بالای آن حرکت می‌کنیم، خوشه‌های با ارزش بیشتر قرار گرفته‌اند. به‌عبارتی، با حرکت به سمت رأس این هرم، با مشترکینی مواجه می‌شویم که ارزش و اهمیت بیشتری از لحاظ مصرف کمتر دارند. برای تعیین مصرف مشترکین در هر خوشه بر مبنای هرم ارزش، از شاخص‌های مصرف مشترکین شامل میزان مصرف، سابقه انشعاب غیرمجاز، تعداد اخطار قطع و زمان پرداخت قبوض برای تعیین ارزش هر مشترک استفاده شده است.

گام هشتم: پیش‌بینی سطح مصرف هر مشترک با استفاده از درخت تصمیم

درخت تصمیم نیز که از سال ۱۹۸۰ مطرح شده، یکی از پرکاربردترین الگوریتم‌های طبقه‌بندی و پیش‌بینی است که یک ساختار درختی شبیه نمودار جریان دارد. مدل‌سازی درخت تصمیم از داده‌های آموزشی در یک جهت تولید خاص بالا به پایین، در دو فاز ساخت و هرس‌کردن[1] انجام می‌شود. این الگوریتم برای ساخت درخت مجموعه داده، به‌طور بازگشتی[2] با توجه به انتخاب ویژگی‌ها به زیرمجموعه‌های کوچکتر بخش‌بندی می‌کند.

۳- یافته‌ها

۳-۱- جمع‌آوری و آماده‌سازی داده‌های پژوهش

شاخص‌ها و معیارهای مؤثر بر خوشه‌بندی مشترکین بر اساس نظر ۱۰ نفر از کارشناسان و خبرگان شرکت آب و فاضلاب شیراز با سابقه بیشتر از ۱۰ سال در ۲ جلسه مشترک با حضور کلیه افراد با استفاده از هم‌اندیشی انتخاب شدند. سپس داده‌های مربوط از پایگاه داده مشترکین شرکت آب و فاضلاب در سال‌های ۱۳۹۸ و ۱۳۹۹ استخراج شد که به شرح جدول ۱ هستند.

به‌منظور آماده‌سازی داده‌های این پژوهش برای ورود به

[1] Purning
[2] Recursively

dx.doi.org/10.22093/wwj.2021.307401.3182

نرم‌افزار:

- به دلیل معکوس بودن داده‌های مشترکین، داده‌ها در بخش تعیین هرم ارزش مصرف مشترکین نرمال‌سازی شدند.
- داده‌های اسمی و عددی، در مرحله اولیه برای ورود به نرم‌افزار کدگذاری شدند.

۳-۲- خوشه‌بندی مشترکین فعلی با الگوریتم خوشه‌بندی

۳-۲-۱- تعیین حداکثر تعداد خوشه با استفاده از شبکه عصبی خودسازمان‌ده

به‌منظور افزایش دقت خوشه‌بندی الگوریتم Kمیانگین، تعداد خوشه ورودی با استفاده از شبکه عصبی خودسازمان‌ده تعیین شد. در واقع این تعداد خوشه به‌عنوان حداکثر تعداد خوشه که الگوریتم Kمیانگین با توجه به آن خوشه‌بندی خواهد شد، در نظر گرفته شد. برای ایجاد الگوریتم از نرم‌افزار SPSS Modeler استفاده شد که با انجام این عملیات تعداد خوشه ۱۰ به‌عنوان خروجی به دست آمد.

۳-۲-۲- اجرای خوشه‌بندی Kمیانگین و تعیین بهترین مدل خوشه‌بندی بر مبنای شاخص دیویس- بولدین

در مرحله قبل، حداکثر تعداد خوشه با استفاده از شبکه عصبی خودسازمان‌ده برابر با ۱۰ مشخص شد؛ در این مرحله الگوریتم K میانگین با تعداد خوشه ۲ تا ۱۰ پیاده‌سازی شد. در ادامه با استفاده از شاخص دیویس- بولدین نتایج خوشه‌بندی ارزیابی شد تا بهترین مدل خوشه‌بندی مشخص شود. برای تعیین تعداد خوشه بهینه، مدل خوشه‌بندی که شاخص دیویس- بولدین در آن بیشترین باشد به‌عنوان بهترین مدل خوشه‌بندی انتخاب شد، بنابراین با توجه به نتایج به‌دست آمده از این شاخص در جدول ۲ مشاهده می‌شود که بهترین خوشه‌بندی متعلق به تعداد خوشه ۶ است.

به این ترتیب عملیات خوشه‌بندی Kمیانگین با تعداد خوشه ۶ انجام شد و در شکل ۲ درصد فراوانی مشترکین در هر خوشه نشان داده شده است.

همان طور که در شکل ۱ مشاهده می‌شود خوشه‌های شماره ۲، ۳ و ۵ بیشترین تعداد مشترکین را شامل شده‌اند. ویژگی‌های هر خوشه در جدول ۳ نشان داده شده است.

خوشه ۱ شامل ۵ درصد از مشترکین است که معمولاً کاربری خانگی داشته و ساکن منطقه ۴ هستند، اکثر آنها مصرف متوسط داشته و سابقه انشعاب غیرمجاز ندارند، عمده مشترکین این خوشه بیشتر از ۵ بار اخطار قطع داشته و زمان پرداخت قبوض آنها بیشتر از ۱۶ روز است.

جدول ۲- شاخص دیویس- بولدین برای تعداد خوشه‌ها

Table 2. Davis-Bouldin index for the number of clusters

Number of clusters	Davis-Bouldin index
2	1.49
3	1.15
4	1.539
5	1.32
6	1.96
7	1.97
8	1.86
9	1.71
10	1.62

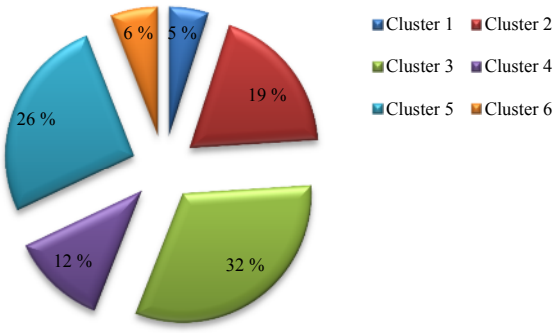

Fig. 2. Percentage of subscribers in each cluster

شکل ۲- درصد فراوانی مشترکین در هر خوشه

خوشه ۲ شامل ۱۹ درصد از مشترکین است که معمولاً کاربری خانگی داشته و ساکن منطقه ۵ هستند، اکثر آنها مصرف کمی داشته و سابقه انشعاب غیرمجاز ندارند، عمده مشترکین این خوشه بین ۲ تا ۴ بار اخطار قطع داشته و زمان پرداخت قبوض آنها بیشتر از ۱۶ روز است.

خوشه ۳ بزرگترین خوشه بوده که شامل ۲۳ درصد از مشترکین است که معمولاً کاربری خانگی و تجاری- صنعتی داشته و ساکن منطقه ۲ هستند، اکثر آنها مصرف زیاد داشته و سابقه انشعاب غیرمجاز ندارند، عمده مشترکین این خوشه بین صفر تا ۱ بار اخطار قطع داشته و زمان پرداخت قبوض آنها کمتر از ۱۵ روز است.

خوشه ۴ شامل ۱۲ درصد از مشترکین است که معمولاً کاربری

جدول ۳- ویژگی مشترکین هر خوشه

Table 3. Characteristics of subscribers of each cluster

Variable / Cluster		5	1	2	3	4	5	6
Number of subscribers (Item)		196300	37750	143450	241600	90600	196300	45300
Type of application (Percent)	Residential	56	91	91	51	80	56	68
	Commercial and Industrial	26	5	5	34	14	26	10
	Public and Administrative	10	1	1	5	2	10	6
	Religious and Educational places	3	1	1	1	1	3	5
	Military	1	1	1	1	1	1	3
	Non-governmental	4	1	1	8	2	4	8
Address (Percent)	Region 1	69	5	5	9	5	69	14
	Region 2	14	8	8	65	5	14	10
	Region 3	5	6	6	12	72	5	6
	Region 4	5	10	10	6	10	5	55
	Region 5	7	71	71	8	8	7	15
Consumption (Percent)	0-12000 (Low consumption)	13	64	64	16	10	13	17
	12,000-15,000 (Medium consumption)	58	28	28	25	82	58	55
	More than 15,000 (High consumption)	29	8	8	59	8	29	28
History of unauthorized branching (Percent)	Yes	6	4	4	1	6	6	7
	No	94	96	96	99	94	94	93
Interrupt warning number (Percent)	From 0 to 1 times (Low)	21	24	24	81	22	21	25
	From 2 to 4 (Medium)	69	68	68	14	24	69	16
	More than 5 times (High)	10	8	8	5	54	10	59
Time to pay bills (Percent)	Less than 15 days (Suitable)	80	13	13	89	12	80	78
	More than 16 days (Inappropriate)	20	87	87	11	88	20	22

خانگی داشته و ساکن منطقه ۳ هستند، اکثر آنها مصرف متوسط داشته و سابقه انشعاب غیرمجاز ندارند، عمده مشترکین این خوشه بیشتر از ۵ بار اخطار قطع داشته و زمان پرداخت قبوض آنها بیشتر از ۱۶ روز است.

خوشه ۵ شامل ۲۶ درصد از مشترکین است که معمولاً کاربری خانگی و تجاری- صنعتی داشته و ساکن منطقه ۱ هستند، اکثر آنها مصرف متوسط داشته و سابقه انشعاب غیرمجاز ندارند، عمده مشترکین این خوشه بین ۲ تا ۴ بار اخطار قطع داشته و زمان پرداخت قبوض آنها بیشتر از ۱۶ روز است.

خوشه ۶ شامل ۶ درصد از مشترکین است که معمولاً کاربری خانگی داشته و ساکن منطقه ۴ هستند، اکثر آنها مصرف متوسط داشته و سابقه انشعاب غیرمجاز ندارند، عمده مشترکین این خوشه

dx.doi.org/10.22093/wwj.2021.307401.3182

بیشتر از ۵ بار اخطار قطع داشته و زمان پرداخت قبوض آنها کمتر از ۱۵ روز است.

۳-۳- تعیین هرم ارزش مصرف مشترکین

برای تعیین مصرف مشترکین در هر خوشه بر مبنای هرم ارزش، ابتدا شاخص‌های مصرف مشترکین شامل میزان مصرف، سابقه انشعاب غیرمجاز، تعداد اخطار قطع و زمان پرداخت قبوض نرمال می‌شود. سپس میانگین مجموع آنها برای مشترکین هر خوشه محاسبه می‌شود. عدد حاصل به‌عنوان ارزش مصرف مشترکین آن خوشه شناخته می‌شود که هرم ارزش مشترکین بر اساس آن ترسیم می‌شود. بر این اساس هرچه عدد ارزش مصرف مشترکین برای یک خوشه بیشتر باشد، آن خوشه شامل مشترکین کم‌مصرف‌تر و با جایگاه بالاتر در هرم ارزش خواهد بود. این موضوع به‌طور شماتیک به شرح شکل ۳ است.

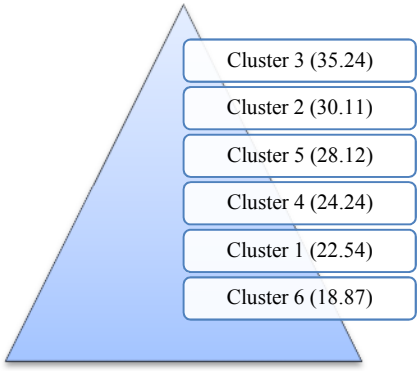

Fig. 3. Pyramid of customer consumption period value

شکل ۳- هرم ارزش دوره مصرف مشترک

در ادامه این ۶ خوشه در ۳ کلاس کلی به‌عنوان مشترکین با مصرف زیاد، متوسط و کم دسته‌بندی شدند. بر این اساس خوشه‌های ۲ و ۳ در کلاس مشترکین با مصرف کم، خوشه‌های ۴ و ۵ در کلاس مشترکین با مصرف متوسط و خوشه‌های ۱ و ۶ در کلاس مشترکین با مصرف زیاد قرار گرفتند. برای پیاده‌سازی درخت تصمیم، کلاس مشترکین به‌عنوان شاخص هدف در نظر گرفته شد.

۳-۴- پیش‌بینی سطح مصرف هر مشترک با درخت تصمیم

در این قسمت به منظور پیش‌بینی سطح مصرف هر مشترک از الگوریتم درخت تصمیم استفاده شد که با توجه به نوع داده‌ها که ترکیبی از داده‌های اسمی و عددی هستند، مناسب‌ترین نوع درخت، درخت تصمیم C۵ است. برای این منظور ۷۰ درصد از کل داده‌ها به‌عنوان داده آموزش و ۳۰ درصد به‌عنوان داده آزمون در نظر گرفته شدند. برای ساخت درخت تصمیم نیز مجموعه داده مشترکین با تعداد ۶ ویژگی شامل نوع کاربری، محل سکونت، میزان مصرف، سابقه انشعاب غیرمجاز، تعداد اخطار قطع و زمان پرداخت قبوض در نظر گرفته شدند. به‌منظور تقسیم مجموعه داده و پیاده‌سازی درخت از نرم‌افزار SPSS Modeler استفاده شد که با ۳ کلاس به‌عنوان ویژگی هدف، صحت درخت ۷۸/۹۲ درصد بود. دقت پیش‌بینی درخت تصمیم برای داده‌های آزمایشی و آزمون در جدول ۴ نشان داده شده است.

جدول ۴- دقت پیش‌بینی درخت تصمیم

Table 4. Accuracy of decision tree

	Experimental set	Test set
True (Percent)	89.25	68.59
False (Percent)	10.75	31.41

از آنجا که هر چه متغیر به ریشه درخت نزدیکتر باشد، تأثیر بیشتری بر هدف دارد، بنابراین می‌توان طبق خروجی‌های نرم‌افزار نتیجه گرفت که از بین ویژگی‌های اولیه به‌عنوان ورودی، به ترتیب زمان پرداخت قبوض، نوع کاربری، تعداد اخطار قطع، محل سکونت و سابقه انشعاب غیرمجاز بیشترین تأثیر را روی متغیر هدف (کلاس مصرف مشترکین) دارند.

الگوریتم درخت تصمیم یک ساختار درخت‌مانند را، به‌عنوان خروجی تولید می‌کند که هر مسیر از گره ریشه به یک گره برگ می‌تواند به‌عنوان یک الگو تفسیر شود. از ریشه تا هر برگ یک الگو، نمایش داده شده است. در واقع قوانین تولید شده به صورت «اگر - آنگاه» هستند و پس از مشاهده خروجی حاصل شده، الگوهای رفتاری استخراج می‌شوند که به شرح جدول ۵ هستند.

جدول ۵- الگوهای استخراج شده از درخت تصمیم‌گیری

Table 5. Patterns extracted from the decision tree

Consumption class	Time to pay bills	Type of application	Interrupt warning number	Address	History of unauthorized branching	Probability (Percent)
Low	Less than 15 days	All	Less than 4 times	Region 5	No	100
	Less than 15 days	Residential	Less than 4 times	Region 2	No	88.23
	It makes no difference	Residential	0 or 1 time	Region 4	No	81.81
	Less than 15 days	Residential-Commercial and Industrial	Less than 4 times	Region 1	No	22.79
Medium	Less than 15 days	Commercial and Industrial	Less than 4 times	Region 4	No	100
	It makes no difference	Commercial and Industrial	Less than 4 times	Region 3	No	77.27
	More than 15 days	All	0 or 1 time	Region 1	No	100
High	More than 15 days	Residential-Commercial and Industrial	More than 5 times	Region 1	No	32.75
	More than 15 days	Commercial and Industrial	More than 5 times	Region 4	No	31.81
	More than 15 days	Residential	More than 5 times	Region 3	No	11.79
	Less than 15 days	Residential-Commercial and Industrial	0 or 1 time	Region 2	No	65.55

همان طور که مشاهده می‌شود طبق خروجی درخت تصمیم، برای تعیین الگوهای مصرفی مشترکین شرکت آب و فاضلاب شیراز ۱۱ الگو استخراج شده که به شرح زیر هستند:

- اگر مشترک هر نوع کاربری داشته و ساکن منطقه ۵ باشد، زمان پرداخت قبوض آن کمتر از ۱۵ روز، تعداد اخطار قطع آن کمتر از ۴ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمال ۱۰۰ درصد یک مشترک کم‌مصرف است.
- اگر مشترک کاربری خانگی و مسکونی داشته و ساکن منطقه ۲ باشد، زمان پرداخت قبوض آن کمتر از ۱۵ روز، تعداد اخطار قطع آن کمتر از ۴ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمال ۸۸/۲۳ درصد یک مشترک کم‌مصرف است.
- اگر مشترک کاربری خانگی و مسکونی داشته و ساکن منطقه ۴ باشد، تعداد اخطار قطع آن صفر یا ۱ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمال ۸۱/۸۱ درصد یک مشترک کم‌مصرف است.
- اگر مشترک کاربری خانگی و تجاری- مسکونی داشته و ساکن منطقه ۱ باشد، زمان پرداخت قبوض آن کمتر از ۱۵ روز، تعداد اخطار قطع آن کمتر از ۴ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمال ۷۹/۲۲ درصد یک مشترک کم‌مصرف است.
- اگر مشترک کاربری تجاری- صنعتی داشته و ساکن منطقه ۴ باشد، زمان پرداخت قبوض آن کمتر از ۱۵ روز، تعداد اخطار قطع آن کمتر از ۴ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمال ۱۰۰ درصد یک مشترک با مصرف متوسط است.
- اگر مشترک کاربری تجاری- صنعتی داشته و ساکن منطقه ۳

dx.doi.org/10.22093/wwj.2021.307401.3182

باشد، تعداد اخطار قطع آن کمتر از ۴ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمای ۷۷/۲۷ درصد یک مشترک با مصرف متوسط است.

- اگر مشترک هر کاربری داشته و ساکن منطقه ۱ باشد، زمان پرداخت قبوض آن بیشتر از ۱۶ روز، تعداد اخطار قطع آن صفر یا ۱ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمای ۱۰۰ درصد یک مشترک با مصرف متوسط است.
- اگر مشترک کاربری خانگی و تجاری- صنعتی داشته و ساکن منطقه ۱ باشد، زمان پرداخت قبوض آن بیشتر از ۱۶ روز، تعداد اخطار قطع آن بیشتر از ۵ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمای ۷۵/۳۲ درصد یک مشترک پرمصرف است.
- اگر مشترک کاربری تجاری- صنعتی داشته و ساکن منطقه ۴ باشد، زمان پرداخت قبوض آن بیشتر از ۱۶ روز، تعداد اخطار قطع آن بیشتر از ۵ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمای ۸۱/۳۱ درصد یک مشترک پرمصرف است.
- اگر مشترک کاربری خانگی داشته و ساکن منطقه ۳ باشد، زمان پرداخت قبوض آن بیشتر از ۱۶ روز، تعداد اخطار قطع آن بیشتر از ۵ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمای ۷۹/۱۱ درصد یک مشترک پرمصرف است.
- اگر مشترک کاربری خانگی و تجاری- صنعتی داشته و ساکن منطقه ۲ باشد، زمان پرداخت قبوض آن کمتر از ۱۵روز، تعداد اخطار قطع آن صفر یا ۱ بار و سابقه انشعاب غیرمجاز نداشته باشد، به احتمای ۶۵/۵۵ درصد یک مشترک پرمصرف است.

۴- بحث و نتیجه‌گیری

شناسایی و پیش‌بینی الگوی رفتاری مصرف مشترکین آب در شرایط بحرانی کنونی، اهمیت و ضرورت زیادی دارد، ازاین‌رو در این پژوهش برای این منظور از تکنیک‌های داده‌کاوی بر مبنای هرم ارزش مصرف مشترکین استفاده شد. مشترکین شرکت آب و فاضلاب شیراز که در این پژوهش بررسی شدند، با توجه به ویژگی‌های میزان مصرف، سابقه انشعاب غیرمجاز، تعداد اخطار قطع و زمان پرداخت قبوض به ۶ خوشه تقسیم شدند که مشترکین هر خوشه ویژگی‌های مشابه با یکدیگر و متفاوت از سایر خوشه‌ها دارند. سپس الگوی رفتاری مصرف مشترکین هر خوشه تعیین شد و میانگین مجموع آنها به‌عنوان الگوی رفتاری مصرف برای کلاسه‌بندی آنها در نظر گرفته شد، به این ترتیب خوشه‌های مشترکین در ۳ کلاس به‌عنوان مشترکین با مصرف کم، مشترکین با مصرف متوسط و مشترکین با مصرف زیاد قرار گرفتند. پس از تفسیر قوانین و ملاحظه روندهای موجود بین الگوهای استخراج شده از درخت تصمیم می‌توان نتیجه گرفت که کلاس مشترکین با مصرف زیاد که ۵۱ درصد از کل مشترکین را شامل شدند. به‌طور عمده کاربری خانگی و تجاری- صنعتی دارند، معمولاً ساکن مناطق ۲ و ۵ بوده و اکثر آنها سابقه انشعاب غیرمجاز ندارند؛ اکثر مشترکین این کلاس، کمتر از ۴ بار اخطار قطع داشته و میانگین زمان پرداخت قبوض آنها حدود ۲۵ روز است. کلاس مشترکین با مصرف متوسط که ۳۸ درصد از کل مشترکین را شامل شده‌اند، به‌طور عمده کاربری خانگی دارند، معمولاً ساکن مناطق ۱ و ۳ بوده و اکثر آنها سابقه انشعاب غیرمجاز ندارند و میانگین زمان پرداخت قبوض آنها حدود ۶۰ روز است. کلاس مشترکین کم‌مصرف که ۱۱ درصد از کل مشترکین را شامل شده‌اند، به‌طور عمده کاربری خانگی دارند، معمولاً ساکن منطقه ۴ بوده و اکثر آنها سابقه انشعاب غیرمجاز ندارند؛ اکثر مشترکین این کلاس، بیشتر از ۵ بار اخطار قطع داشته و میانگین زمان پرداخت قبوض آنها حدود ۱۰۹ روز است.

در ادامه برمبنای این ۳ کلاس، ۱۱ الگوی رفتاری برای مصرف مشترکین شرکت آب و فاضلاب شیراز استخراج شد که بر اساس آنها می‌توان میزان مصرف مشترکین جدید را پیش‌بینی کرد. به این ترتیب که با ورود هر مشترک جدید، کلاس ارزشی مشترک از نظر نوع مصرف پیش‌بینی می‌شود، بر این اساس می‌توان تشویق‌هایی برای مشترکین کم‌مصرف که در جایگاه بالای هرم ارزش قرار گرفته‌اند، تدارک دید و برنامه‌هایی برای هشدار به مشترکین پرمصرف در راستای کاهش مصرف آب و آشناسازی آنها با روش‌های صحیح مصرف در نظر گرفت.

مهمترین محدودیت انجام این پژوهش، دشواری در جمع‌آوری داده‌های مشترکین شرکت آب و فاضلاب شیراز بود که یک پایگاه داده یکپارچه نداشته و اطلاعات مشترکین از دو پایگاه مجزا جمع‌آوری و یکپارچه‌سازی شدند.

در ادامه پیشنهاداتی برای پژوهشگران آینده ارائه می‌شود:

- به‌منظور استخراج الگوهای رفتاری به شکل «اگر-آنگاه» پیشنهاد می‌شود از تکنیک قوانین انجمنی نیز در کنار درخت تصمیم استفاده

شود. مزیت این روش، آن است که ارتباط بین شاخص‌ها را نیز بررسی کرده و قوانینی در این رابطه نیز استخراج می‌کند.

• با توجه به اهمیت انشعاب‌های غیرمجاز و تأثیر آنها در میزان مصرف آب، پیشنهاد می‌شود در پژوهشی به بررسی و شناسایی عوامل مؤثر بر ایجاد این نوع انشعاب‌ها پرداخته شود.

۵- قدردانی

این پژوهش بر مبنای آمار و اطلاعات مشترکین شرکت آب و فاضلاب شیراز انجام شد. از مجموعه نامبرده به دلیل تأمین اطلاعات و تجهیزات موردنیاز قدردانی می‌شود.

References

Aghababaei, A. & Shahrabi, J. 2012. Application of data mining knowledge in the face of indebted subscribers in Mashhad water and wastewater company (Region 3). *The 6th Data Mining Conference*. Tehran, Iran. (In Persian)

Aghababaei, A., Shahrabi, J. & Hadavandi, I. 2011. Designing a data mining model for indebted subscribers of Mashhad water and wastewater company (Region 3). *5th Iran Data Mining Conference,* Amirkabir University of Technology. Tehran, Iran. (In Persian)

Aghahosseinali Shirazi, M., & Akbarpour, A. 2011. Estimation of daily urban water demand using the February series: a case study of Birjand city in South Khorasan province. *International Conference on Water and Wastewater*. Tehran, Iran. (In Persian)

Ahangarkani, M. & Khasteh, S. H. 2019. Analysis of urban (domestic) water consumption in Babol city using data mining methods. *Sepehr Geographical Information*, 28**,** 53-69. (In Persian)

Amini, Q. 2020. Modeling the diagnosis of unauthorized water use (case study: Qom city). *Journal of Water and Wastewater*, 31(4)**,** 184-193. (In Persian)

Amini, Q., Farmani Enteza, H., Jan Sadeghpour, A. & Davoodabadi, A. 2018a. Application of data mining in identifying subscribers with unauthorized water uses (case study: Qom water and wastewater company). *2nd Iranian Water and Wastewater Science and Engineering Congress*. Isfahan, Iran. (In Persian)

Amini, Q., Farmani Entezam, H., Jan Sadeghpour, A. & Davoodabadi, A. 2018b. Identification and extraction of water consumption pattern by data mining method (case study: Qom water and wastewater company). *2nd Iranian Water and Wastewater Science and Engineering Congress*. Isfahan, Iran. (In Persian)

Amoozegar, M. 2016. Presenting a two-step solution to identify the pattern of electricity consumption. *Iranian Journal of Quality and Productivity of Electricity Industry*, 5(9)**,** 48-57. (In Persian)

Ansari, H., Boostani, A., Tabatabayi, A. & Foroozesh. M. 2015. Investigation of consumption management and estimation of Mashhad drinking water demand in the horizon of 1420. *Water and Sustainable Development*, 4(1)**,** 125-132. (In Persian)

Arora, N., Arora, A. S., Sharma, S. & Reddy, A. S. 2014. Use of cluster analysis-A data mining tool for improved water quality monitoring of river Satluj. *International Journal of Advanced Computer Science and Applications*, 6**,** 63-69.

Azhar, S. A. S., Johar, H., Baki, S. R. M. S. & Tahir, N. M. 2013. Optimization of water quality monitoring based on fuzzy algorithms. *In 2013 IEEE Conference on Systems, Process and Control (ICSPC)*. Kuala Lumpur, Malaysia. 283-288.

Azimi, V., Vakilifard, A. & Asadi, A. 2015. Evaluation of M5 gene expression planning and model in estimating daily flows, case study of Liquan river. *International Quarterly Journal of Water Resources Analysis and Development*, 3**,** 134-142.

Berry, M. J. & Linoff, G. S. 2004. *Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management*, John Wiley and Sons, Indiana, USA.

Boyle, C. E., Eskaf, S., Tiger, M. W. & Hughes, J. A. 2011. Mining water billing data to inform policy and communication strategies. *Journal-American Water Works Association*, 103**,** 45-58.

Castellano, I. M. 2020. *Water Scarcity in the American West*, Palgrave Macmilan, Cham. New York, USA. 51-93.

Chen, X., Yang, S. H., Yang, L. & Chen, X. 2015. A benchmarking model for household water consumption based on adaptive logic networks. *Procedia Engineering*, 119**,** 191-198.

Cho, Y. 2016. A watershed water quality evaluation model using data mining as an alternative to physical watershed models. *Water Science and Technology: Water Supply*, 16**,** 703-714.

Davidson, I. 2002. *Understanding K-Means Non-Hierarchical Clustering*. Suny Albany, Technical Report, 02-2.

Davies, D. L. & Bouldin, D. W. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2**,** 224-227.

Díaz, J. L., Herrera, M., Izquierdo, J. & Pérez-García, R. 2010. The tasks of pre and post-processing in data mining applied to a real world problem. *5th International Congress on Environmental Modelling and Software,* Ottawa, Ontario, Canada.

Dutta, P. & Chaki, R. 2012. A survey of data mining applications in water quality management. *In Proceedings of the CUBE International Information Technology Conference*, Pune, India.

Ebrahimi, P. & Naderi, H. 2001. Study and evaluation of drinking water supply and demand management in drought conditions of Isfahan. *Journal of Water and Environment*, 48-49**,** 89-97. (In Persian)

Eskandary, M., Taghavifard, M. T., Raeesi Vanani, I. & Ghazi Noori, S. 2020. Identification and prioritization of public-private partnership indicators in Iran's water and wastewater industry via data mining algorithms. *Iranian Journal of Economic Studies*, 8(2)**,** 375-396.

Foster, H. S. & Beattie, B. R. 1981. On the specification of price in studies of consumer demand under block price scheduling. *Land Economics*, 57(4)**,** 624-629.

Garcia, X., Ribas, A. Llausàs, A. & Saurí, D. 2013. Socio-demographic profiles in suburban developments: implications for water-related attitudes and behaviors along the Mediterranean coast. *Applied Geography*, 41**,** 46-54.

Gu, Q., Deng, J., Wang, K., Lin, Y., Li, J., Gan, M., et al. 2014. Identification and assessment of potential water quality impact factors for drinking-water reservoirs. *International Journal of Environmental Research and Public Health*, 11**,** 6069-6084.

Humaid, E. H. 2012. A data mining based fraud detection model for water consumption billing system in MOG. MSc. Thesis. Islamic University of Gaza Deanery of Higher Studies Information Technology Program Department of Computer Science. Gaza, Palestine.

Jahanpour, K. & Nosrati, G. 2015. An overview of data mining applications in water and wastewater company management; methods and patterns. *The 1st International Conference on Humanities with Indigenous-Islamic Approach and Emphasis on New Research*. Sari, Iran. (In Persian)

Javadianzade, M. M. 2009. Preparation of urban water demand function using artificial neural networks method in Yazd. *3rd National Conference on Water and Wastewater with Consumption Pattern Modification Approach*. Niroo Research Institute. Tehran, Iran. (In Persian)

Ji, Y., Lei, X., Cai, S. & Wang, X. 2016. Application of a classifier based on data mining techniques in water supply operation. *Water*, 8**,** 599.

Kazemi, Z. 2015. Applying process mining to improve knowledge management processes in contact centers (case study: contact center 122 of Tehran water and wastewater organization). MSc. Thesis, Tarbiat Modares University. Tehran, Iran. (In Persian)

Khalfi, A. A., Shiri, Q. & Pourashraf, Y. 2018. Investigating the pattern of domestic water consumption with the approach of consumer segmentation (case study: household water consumers in Ilam city). *Journal of Water and Wastewater*, 29(2)**,** 59-67. (In Persian)

Khan, M. A., Islam, M. Z. & Hafeez, M. 2012. Evaluating the performance of several data mining methods for predicting irrigation water requirement. *In Proceeding of the 10th Australasian Data Mining Conference*, 134**,** 199-208.

Kojury Naft Chali, M. & Fereydonian, A. 2015. Identifying the pattern of electricity consumption by data mining. *30th International Conference on Electricity*. Tehran, Iran. (In Persian)

Kovács, F., Legány, C. & Babos, A. 2013. *Cluster Validity Measurement Techniques, Department of Automation and Applied Informatics*. Budapest University of Technology and Economics, Budapest, Hungary.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, California, Los Angeles. USA. 1**,** 281-297.

Maleki Nasab, A., Abrishamchi, A. & Tajrishi, M. 2007. Evaluation of household water consumption savings through the use of reducing components. *Journal of Water and Wastewater*, 18(2)**,** 2-11. (In Persian)

Malekmohamadi, M. & Mozafari, M. M. 2018. Applying social marketing in optimize management of water consume. *Quarterly Journal of Interdisciplinary Studies in the Humanities,* 10(4)**,** 197-218. (In Persian)

Mohammadi, R. 2014. Investigating the effect of targeted subsidies on consumption pattern and water consumption in Ardabil. MSc. Thesis, Islamic Azad University Garmi Branch, Ardabil, Iran. (In Persian)

Mohammad Taheri, A. 2013. Prediction of effluent quality of wastewater treatment plant using predictive data mining-case study: Baharan industrial town. MSc. Thesis, Hamadan, Iran. (In Persian)

Monedero, I., Biscarri, F., Guerrero, J., Roldan, M. & Leon, C. 2015. An approach to detection of tampering in water meters. *19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems*. National Uiversity of Singapore, Singapore.

Monika, C. & Amarpreet, K. 2018. A comparative study of classification techniques for fraud detection. *Journal on Future Revolution in Computer Science and Communication Engineering*, 4**,** 19-23.

Noori, A., Banihabib, M. E. & Soltani, J. 2015. Determining and prioritization of sustainable management strategies for water supply and consumption in the dry areas of Iran. *The 1st Conference and Exhibition of Water Science and Engineering*. Shahid Beheshti University. Terhan, Iran. (In Persian)

Oyedepo, S. O. 2014. Towards achieving energy for sustainable development in Nigeria. *Renewable and Sustainable Energy Reviews*, 34**,** 255-272.

Rathnayaka, K., Maheepala, S., Nawarathna, B., George, B., Malano, H., Arora, M., et al. 2014. Factors affecting the variability of household water use in Melbourne, Australia. *Resources, Conservation and Recycling*, 92**,** 85-94.

Rayegan Shirazinejad, A., Zare, M., Zare, F., Banshi, M. M. & Rezaei, S. 2015. Investigation of statistical models in modeling wastewater treatment process using data mining method. *Environmental Health Engineering*, 2**,** 186-194. (In Persian)

Rosenblatt, F. 1962. *Principles of Neurodynamics Spartan*. New York, USA.

Sabouhi, M. & Noubakht, M. 2009. Estimating the water demand function of Pardis city. *Journal of Water and Wastewater*, 20(2)**,** 69-74. (In Persian)

Sattari, M. T. & Rezazadeh Judi, A. 2018. Monthly runoff modeling using data mining methods based on feature selection algorithms. *Protection of Water and Soil Resources*, 7**,** 39-53. (In Persian)

Sattari, M. T., Abbasgholi Nayebzadeh, M. & Mirabbasi Najafabadi, R. 2014. Predicting surface water quality using decision tree method. *Irrigation and Water*, 4**,** 76-88. (In Persian)

Shamsai, M. 2000. Estimating the water demand function of Isfahan province. a collection of 21 papers presented in the first scientific and research conference on water consumption optimization. *Public Relations and International Affairs Publications of Tehran Water and Wastewater Company*. Terhan, Iran. (In Persian)

Soleimanpour, S. M., Hedayati, B. & Zolfaghari, M. 2015. Determining the effective indicators on drinking water quality using QUEST data mining technique in Saadatshahr, Fars province. *3rd International Conference on Rainwater System*. Birjand, Iran. (In Persian)

Soleimanpour, S. M., Mesbah, S. H. & Hedayati, B. 2016. Application of K-Means and CART data mining algorithms in determining the most effective factors of drinking water quality in Noorabad plain of Fars province. *11th National Conference on Watershed Management Science and Engineering*. Yasouj, Iran. (In Persian)

Soleimanpour, S. M., Mesbah, S. H. & Hedayati, B. 2018. Application of CART decision tree data mining techniques in determining the most effective drinking water quality factors (case study: Kazerun plain, Fars province). *Health and Environment*, 11(1)**,** 1-14. (In Persian)

Tabatabai, A. 2009. *An Attitude on Data Mining*. Qazvin Azad University Pub., Qazvin, Iran. (In Persian)

Tabesh, M. & Dini, M. 2010. Predicting daily urban water demand using artificial neural networks, case study: Tehran. *Journal of Water and Wastewater*, 21**,** 84-95. (In Persian)

Taherdoost, M. A. 2021. Prediction of domestic drinking water demand in Shiraz city using time series and panel data. *The 1st National Conference on Water Quality Management and the 3rd National Conference on Water Consumption Management with the Approach of Reducing Waste and Recycling*. Tehran University. Tehran, Iran. (In Persian)

Thompson, E. 2016. Investigating drinking water advisories in first nations communities through data mining. MSc. Thesis. University of Guelph. Hamilton, Canada.

Wen, Y. Y., Huang, W. M., Wu, J., Chen, Y. & Song, J. Q. 2013. Water consumption analysis system based on data mining. *Applied Mechanics and Materials*, 241**,** 1093-1097.

Willis, R. M., Stewart, R. A., Giurco, D. P., Talebpour, M. R. & Mousavinejad, A. 2013. End use water consumption in households: impact of socio-demographic factors and efficient devices. *Journal of Cleaner Production*, 60**,** 107-115.

Yurekli, K., Taghi Sattari, M., Anil, A. & Hinis, M. 2012. Seasonal and annual regional drought prediction by using data-mining approach. *Atmósfera*, 25**,** 85-105.